International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 1, February 2021, pp. 620~627
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i1.pp620-627

A space-structure based dissimilarity measure for categorical data

Kevin Alejandro Hernández, D. Cárdenas Peña, and Álvaro A. Orozco
Automatics Reseach Group, Faculty of Engineering, Universidad Tecnológica de Pereira (UTP), Colombia

Article Info

Article history:

Received Feb 5, 2020
Revised Jul 9, 2020
Accepted Jul 20, 2020

Keywords:

Categorical data
Clustering
Dissimilarity
Similarity
Space-structure

ABSTRACT

The development of analysis methods for categorical data begun in 90's decade, and it has been booming in the last years. On the other hand, the performance of many of these methods depends on the used metric. Therefore, determining a dissimilarity measure for categorical data is one of the most attractive and recent challenges in data mining problems. However, several similarity/dissimilarity measures proposed in the literature have drawbacks due to high computational cost, or poor performance. For this reason, we propose a new distance metric for categorical data. We call it: weighted pairing (W-P) based on feature space-structure, where the weights are understood like a degree of contribution of an attribute to the compact cluster structure. The performance of W-P metric was evaluated in the unsupervised learning framework in terms of cluster quality index. We test the W-P in six real categorical datasets downloaded from the public UCI repository, and we make a comparison with the distance metric (DM3) method and hamming metric (H-SBI). Results show that our proposal outperforms DM3 and H-SBI in different experimental configurations. Also, the W-P achieves highest rand index values and a better clustering discriminant than the other methods.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Kevin Alejandro Hernández,
Faculty of Engineering,
Universidad Tecnológica de Pereira (UTP),
Pereira, Colombia.
Email: kevin_loco@utp.edu.co

1. INTRODUCTION

The augment of available datasets provides to the research community new resources to achieve scientific discoveries, optimizing industrial processes and it grants to find relations or characteristic patterns in data [1]. However, there are open issues, for example, determining a dissimilarity measure is one of the most attractive and recent challenges in data mining problems. This is because the performance of many algorithms for clustering, classification, dimensionality reduction, and outliers detection, depends on the metric used to measure similarity/dissimilarity among the data [2]. For this reason, it is convenient to establish an appropriate distance measure for a given data set, instead of using an arbitrary metric.

Choosing a metric for quantitative data (continuous) is relatively simple, since there are several developed metrics such as Euclidean, Cosine, Manhattan, among others. Also, with this type of data it can be used the standard methods of machine learning directly and performing numeric calculations without drawbacks or limitations [3, 4]. While, choosing a metric for categorical data (or nominal) is more complex, due to there is not an intrinsic similarity/dissimilarity measure established for categorical objects [5]. In addition, standard machine learning algorithms can not be applied directly in categorical data, because it is

not appropriate to calculate statistic descriptors (mean, standard deviation, etc) over a dataset with nominal or qualitative variables as if they were quantitative variables [6]. Besides, categorical data is highly overlapped.

This observation has motivated several researchers to work with categorical data, as in the case of [7] who carried out a study on distances for heterogeneous data (databases with mixed quantitative and qualitative variables), based on a supervised learning approach where each sample has additional information about the class to which it belongs. However, this approach can be extended to the unsupervised learning paradigm. Other works such as [8-12] was proposed to binarize the categorical information, resulting in samples that are assigned a 1 or a 0 as indicated by their original qualitative value, to later using a similarity/dissimilarity measures for binary data in cluster algorithms. Nevertheless, far from being a reliable solution these methods have a problem: they are only applicable to categorical databases whose variables have only two possible states, which in general is not the case. Further, these algorithms need to handle a large number of binary attributes when the datasets have features with many more categories, classes or groups. The above in increases the computational cost and memory storage of the algorithm [13].

Alternatively, in the work developed by [14], the authors evaluate the performance of a variety of similarity measures in the literature like overlap, inverse occurrence frequency, occurrence frequency, among others. This is done in the context of outliers detection, but their experiments showed a very poor performance and unstable results with increased standard deviation, and suggest that there is no one best performing similarity measure, and it is necessary to understand how a similarity measure handles the different attributes of categorical datasets [15]. In a recent research [16] was proposed a new metric to measure the distance between categorical type objects, based on the frequency probability of each attribute value in the whole dataset and the degree of dependence among different attributes. However, the probability distribution of the attributes must be taken into account, and the inherent structure of the data in the feature space is not considered [14, 17].

Given the previously pointed out, in this paper, we propose a new approach to determine the similarity/dissimilarity measure between qualitative data based on the number of possible states of a categorical variable, to assign the degree of relevance or degree of contribution to the compact cluster structure of each attribute. We call our method: weighted pairing (W-P) based on feature space-structure. The effectiveness of the proposed W-P metric is demonstrated by performing experiments in real categorical databases obtained from the public UCI machine learning repository [18]. We compare our proposal with the distance metric method (DM3) and hamming (H-SBI). We analyze the performance of the W-P distance metric by embedding it into the framework of the K-modes algorithm, which is the most popular distance-based clustering method for purely categorical data [16, 19], using a centroid initialization method proposed in [20].

2. WEIGHTED PAIRING DISTANCE (W-P)

In this section, we introduce the dissimilarity measure for categorical data in the paradigm of unsupervised learning as a weighted pairing distance learned according to the attribute compactness within the data structure. Let a set of N objects $\mathcal{X} = \{\boldsymbol{x}_n : n \in [1, N]\}$, each of them expressed as the vector $\boldsymbol{x}_n = [x_{n1}, \dots x_{np}, \dots x_{nP}]$, where P stands for the number of attributes. For categorical data, the object n at attribute p, $x_{np} \in \mathcal{A}_p$, takes one value from the unordered, discrete set $\mathcal{C}_p = \{c_{pd} : d \in [1, D_p]\}$ of D_p possible values [21]. In order to establish a similarity measure between categorical objects, the simple matching function aggregates the number matching values [3]:

$$v(\boldsymbol{x}_n, \boldsymbol{x}_m) = \frac{1}{P}\sum_{p=1}^{P} \delta(x_{np}, x_{mp}) \quad (1)$$

being $\delta(x_{np}, x_{mp})$ the delta function that equals to 1 if $x_{np} = x_{mp}$ or 0 in otherwise. Equation (1) determines how many attribute values a couple of samples have in common, and $1/P$ is a normalization factor. However, the simple matching lacks of an attribute ranking required for understanding the categorical data [15, 22].

Aiming to overcome above issue, we propose a dissimilarity measure that considers the relevance of each attribute, termed weighted pairing (W-P) distance, as follows:

$$d_{\boldsymbol{w}}(\boldsymbol{x}_n, \boldsymbol{x}_m) = 1 - \sum_{p=1}^{P} w_p \delta(x_{np}, x_{mp}) \quad (2)$$

$$s.t. \ \sum_{p=1}^{P} w_p = 1 \quad (3)$$

with the normalized relevance weights $w_p \in [0, 1]$ satisfying $\sum_{p=1}^{P} w_p = 1$, and $d_{\boldsymbol{w}}(\boldsymbol{x}_n, \boldsymbol{x}_m) \in [0, 1]$.

To determine the attribute relevance, we assume that the more possible values an attribute can take, the more dispersed the objects in the feature space, as Figure 1 illustrates. Then, we account for the data compactness in the relevance weights in terms of the attribute cardinality as:

$$w_p = \frac{|\mathcal{C}_p|^{-1}}{\sum_{q=1}^{P}|\mathcal{C}_q|^{-1}} \tag{4}$$

where operator $|\cdot|$ determines the number of objects within a set. Therefore, the most relevant attributes are those contributing to the most to data compactness, since the attribute relevance becomes inversely proportional to its number of possible values.

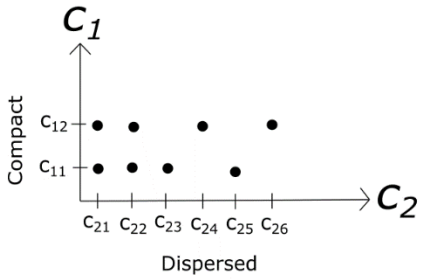

Figure 1. Attribute space example for categorical data, with $L_1 = 2$ and $L_2 = 6$

2.1. Weighted pairing distance properties

If the attribute weights satisfy (3), the dissimilarity function in (2) becomes a distance.

Lemma 1. $d_{\boldsymbol{w}}(\boldsymbol{x}_n, \boldsymbol{x}_m)$ *is non − negartive.*

Proof. Let $d_{\boldsymbol{w}}(\boldsymbol{x}_n, \boldsymbol{x}_m) \geq 0$:

$$1 - \sum_{p=1}^{P} w_p \delta(x_{np}, x_{mp}) \geq 0$$

$$\sum_{p=1}^{P} w_p \delta(x_{np}, x_{mp}) \geq 1$$

$$\max \sum_{p=1}^{P} w_p \delta(x_{np}, x_{mp}) = 1$$

which holds if $\delta(\cdot,\cdot) \leq 1$, and w_p satisfies (3).

Lemma 2. $d_{\boldsymbol{w}}(\boldsymbol{x}_n, \boldsymbol{x}_m)$ *is symmetric.*

Proof. Let $d_{\boldsymbol{w}}(\boldsymbol{x}_n, \boldsymbol{x}_m) = d_{\boldsymbol{w}}(\boldsymbol{x}_n, \boldsymbol{x}_m)$:

$$1 - \sum_{p=1}^{P} w_p \delta(x_{np}, x_{mp}) = 1 - \sum_{p=1}^{P} w_p \delta(x_{np}, x_{mp})$$

$$\sum_{p=1}^{P} w_p \delta(x_{np}, x_{mp}) = \sum_{p=1}^{P} w_p \delta(x_{np}, x_{mp})$$

due to $\delta(x_{np}, x_{mp}) = \delta(x_{np}, x_{mp})$ the property is satisfied.

Lemma 3. $d_{\boldsymbol{w}}(\boldsymbol{x}_n, \boldsymbol{x}_m) = 0$ *if and only if* $\boldsymbol{x}_n = \boldsymbol{x}_m$.

Proof. Let $d_{\boldsymbol{w}}(\boldsymbol{x}_n, \boldsymbol{x}_m) = 0$:

$$1 - \sum_{p=1}^{P} w_p \delta(x_{np}, x_{mp}) = 0$$

$$\sum_{p=1}^{P} w_p \delta(x_{np}, x_{mp}) = 1$$

which is the maximum value that the linear combination takes according to Lemma 1. Such a maximum is achieved if and only if $w_p \geq 0$ satistifes Equation (3) and $\delta(x_{np}, x_{mp}) = 1$ for all p, that is, $\boldsymbol{x}_n = \boldsymbol{x}_m$.

Lemma 4. $d_{\boldsymbol{w}}(\boldsymbol{x}_n, \boldsymbol{x}_m)$ *satisfies the triangle inequality.*

Proof. The triangle inequality states that for any three objects $\boldsymbol{x}_n$, $\boldsymbol{x}_m$, $\boldsymbol{x}_o$, it must be satisfied that $d_{\boldsymbol{w}}(\boldsymbol{x}_n, \boldsymbol{x}_m) \leq d_{\boldsymbol{w}}(\boldsymbol{x}_n, \boldsymbol{x}_o) + d_{\boldsymbol{w}}(\boldsymbol{x}_o, \boldsymbol{x}_m)$. Then:

$$1 - \sum_{p=1}^{P} w_p \delta(x_{np}, x_{mp}) \leq 1 - \sum_{p=1}^{P} w_p \delta(x_{np}, x_{op}) + 1 - \sum_{p=1}^{P} w_p \delta(x_{op}, x_{mp})$$

$$\sum_{p=1}^{P} w_p \left(\delta(x_{np}, x_{op}) + \delta(x_{op}, x_{mp}) - \delta(x_{np}, x_{mp}) \right) \leq 1 \quad (5)$$

To prove by contradiction, (5) is not satisfied when $\delta(x_{np}, x_{op}) + \delta(x_{op}, x_{mp}) - \delta(x_{np}, x_{mp}) > 1$ for any p, that is

$$\delta(x_{np}, x_{op}) = \delta(x_{op}, x_{mp}) = 1 \quad (6)$$

$$\delta(x_{np}, x_{mp}) = 0 \quad (7)$$

Equation (6) implies $x_{np} = x_{op}$ and $x_{op} = x_{mp}$. Therefore, $x_{np} = x_{op} = x_{mp}$ yields $\delta(x_{np}, x_{mp}) = 1$, which by contradicting (7), proves (5).

2.2. Distance implementation

The computation of the proposed W-P distance is a two-step process. Firstly, we need to learn the weights w_p given a dataset as shown in Algorithm 1. Secondly, given the weights and an object pair, we compute the W-P distance following Algorithm 2.

Algorithm 1 Distance weights computation

```
function WP WEIGHT(Dataset 𝒳)
    for p = 1 to P do
        L_i = |C_p|
    end for
    L = 1 / (Σ_{p=1}^{P} L_p^{-1})
    for p = 1 to P do
        w_p = L / L_i
    end for
end function
```

Algorithm 2 WP dissimilarity measure

```
function WP METRIC(w, Dataset 𝒳, μ )
    for n = 1 to N do
        for k = 1 to K do
            for p = 1 to P do
                if x_np = μ_kp then
                    δ = 1
                else  δ = 0
                end if
            end for
            d_w = 1 − w.δ
        end for
    end for
end function
```

where $\boldsymbol{x}_n$ is a 1-by-P vector containing a single observation. $\boldsymbol{\mu}$ is an K-by-P matrix containing all centroids, $d_{\boldsymbol{w}}$ is an N-by-K matrix of distances between all observations $\boldsymbol{x}$ and all centroids $\boldsymbol{\mu}$.

According to Algorithm 1, W-P demands the computation of P cardinality values and a scaling factor L, that yields a time cost of $\sim\mathcal{O}(4P)$. Besides, Algorithm 2 verifies an attribute matching P times for NK object pair with a complexity of $\sim\mathcal{O}(3NKP)$. Therefore, the computational complexity of W-P is linear on the number of attributes, $\mathcal{O}(2P)$.

3. EXPERIMENTAL SETUP

We analyze the performance of the W-P distance in clustering tasks by embedding it into the K-modes algorithm [23] and initialized cluster centroids with [20] method. Table 1 summarizes the details of the six UCI machine learning datasets considered for evaluating the proposed distance [18]. The first dataset, *Congressional Voting Records*, includes 16 key votes identified by the Congressional Quarterly Almanac of the US House of Representatives, grouped into Democrat or Republican. The *Breast Cancer Wisconsin (Original)* collection holds 699 samples periodically collected from 1989 to 1991 as the clinical cases of Dr. Wolberg and labeled as benign or malignant. The *Mushroom* dataset corresponds to 23 species of gilled mushrooms in the Agaricus and Lepiota families. The dataset describes each specimen in terms of its physical characteristics and classifies it as poisonous or edible. *Soybean (Small)* contains 35 categorical attributes, among nominal and ordered, and four disease classes. The *Car Evaluation*, a dataset derived from a hierarchical decision model, labels 1728 cars according to six different aspects into unacceptable, acceptable good, and very good quality. Lastly, the *Zoo* dataset contains sixteen boolean-valued attributes to group specimens into seven different animal classes.

Table 1. Description of UCI public datasets

Data sets	# of simples	# of features	Classes
Voting	435	16	2
WBCD	699	9	2
Mushroom	8124	22	2
Soybean	47	35	4
Car	1728	6	4
Zoo	101	16	7

For each considered dataset, we assess the clustering performance of W-P in a 50 fold bootstrap scheme, and report average and standard deviation for the following cluster quality indexes, the average intracluster/intercluester distance, cluster discrimination index, rand index and normalized mutual information index. We compare our proposal with the state-of-the-art methods, distance metric (DM3) and Hamming metric with support based initialization (H-SBI), methods proposed in [16] and [20] respectively.

Cluster discrimination index (CDI): Given K clusters $\mathcal{X}_k \subset \mathcal{X}$ with $\bigcup_k \mathcal{X}_k$ and $\mathcal{X}_k \cap \mathcal{X}_{k'} = \emptyset$, the CDI computes the performance according to the average intracluster distances (AID) as [16]:

$$CDI = \frac{1}{K}\sum_{k=1}^{K} \frac{AID(\mathcal{X}_k,\mathcal{X}_k)}{\frac{1}{K-1}\sum_{k'\neq k} AID(\mathcal{X}_k,\mathcal{X}_{k'})}$$

$$AID(\mathcal{X}_k,\mathcal{X}_{k'}) = \frac{\sum_{\boldsymbol{x}_n\in\mathcal{X}_k}\sum_{\boldsymbol{x}_m\in\mathcal{X}_{k'}} d_{\boldsymbol{W}}(\boldsymbol{x}_n,\boldsymbol{x}_m)}{N_k N_{k'}}$$

where N_k stands for the cardinality of the k-th cluster. Therefore, $CDI \geq 0$, and the smaller its value, the more distant the clusters and the closer the objects within each cluster. Where $\boldsymbol{\mu}_k$ denotes the centroid of the k-th cluster resulting from the K-modes algorithm, and Δ_k stands for the average distance between $\boldsymbol{\mu}_k$ and all the k-th cluster objects.

Rand index (RI): is a similarity measure based on the overlap in class agreement, compared to the class disagreement, is defined as [24]:

$$RI = \frac{Tr\{CM\}}{N}$$

being $Tr\{\cdot\}$ the trace operator and $CM \in [0,N]^{K\times C}$ corresponds to the permuted confusion matrix between the cluster algorithm output l_k and *gold standard* label y_c.

Normalized mutual information (NMI): The NMI score relies on the shared object membership, is a symmetric measure for the degree of dependency between l_k and y_c. Unlike correlation, mutual information also takes higher order dependencies into account [25]:

$$NMI = \frac{\sum_{k=1}^{K}\sum_{c=1}^{C} N_{kc}\log\left(N\frac{N_{kc}}{N_k N_c}\right)}{\sqrt{\left(\sum_{k=1}^{K} N_k \log\left(\frac{N_k}{N}\right)\right)\left(\sum_{c=1}^{C} N_c \log\left(\frac{N_c}{N}\right)\right)}}$$

where N_c, N_k, N_{kc} denote the number of objects labeled as $y_n = c$, the number of objects in the k-th cluster, and the number of objects grouped in the k-th cluster that belong to c-th ground truth label, respectively. NMI is a positive value with a maximum of 1 achieved when the ground-truth and the resulting clustering perfectly match.

4. RESULTS AND DISCUSSION

In order to determine the best metric performance between DM3, H-SBI and W-P, we validate our proposal metric distance in terms of intracluster/intercluster distance, CDI, RI and NMI index in the unsupervised learning framework. The first evaluated index was CDI, a smaller CDI value indicate better discrimination on the cluster structure of the dataset. In the Table 2 we compare our W-P metric with DM3 and H-SBI. We see that for three of the six datasets we obtain better CDI values, this may be a consequence of the allocation of the weights, giving greater relevance to the attributes that make the structure of the data more compact.

Table 2. CDI obtained by the different metrics on six real data sets

Data sets	W-P	DM3	H-SBI
Voting	**0.4286**	0.4342	0.4459
WBCD	0.7330	**0.3374**	0.7787
Mushroom	0.6991	0.7150	**0.6111**
Soybean	0.2826	**0.2086**	0.3220
Car	**0.7895**	0.7919	0.8152
Zoo	**0.2024**	0.2656	0.2112

As we can see that for distance-based clustering on categorical data, the K-modes algorithm with the proposed distance metric has a competitive advantage in terms of clustering rand index, in Table 3 the distance W-P metric obtain a betters results in four of six datasets in comparison with DM3 and H-SBI, specifically in the Voting and Mushroom datasets the RI index increase drastically. In addition to making an exhaustive evaluation of the W-P metric introduced in the K-modes clustering algorithm with support based initialization, the NMI index was evaluated. And as can see in Table 4 we obtained bets results to DM3 and H-SBI, we exceeded them in four of the six datasets, the results indicate that our proposed distance metric is more appropriate for the unsupervised categorical data analysis.

Table 3. Clustering performance in terms of RI of K-modes algortihm with the different distance metrics

Data sets	W-P	DM3	H-SBI
Voting	**0.8992±0.0056**	0.7823±0.0016	0.8639±0.0181
WBCD	**0.9192±0.0130**	0.8827±0.0752	0.4996±0.0737
Mushroom	**0.7662±0.0156**	0.6732±0.0880	0.6573±0.0940
Soybean	**0.9861±0.0284**	0.9314±0.0758	0.7800±0.1769
Car	0.4366±0.0148	**0.5059±0.0123**	0.1670±0.1350
Zoo	0.7424±0.0425	**0.9064±0.0450**	0.5224±0.1395

Table 4. Clustering performance in terms of NMI of K-modes algorithm with the different distance metrics

Data sets	W-P	DM3	H-SBI
Voting	**0.5546±0.0242**	0.4987±0.0078	0.4542±0.0450
WBCD	0.6049±0.0445	**0.6917±0.1304**	0.0745±0.0724
Mushroom	0.2189±0.0292	**0.3182±0.1372**	0.1099±0.1060
Soybean	**0.9741±0.0517**	0.8991±0.1089	0.8030±0.1459
Car	**0.1145±0.0320**	0.0725±0.0253	0.0483±0.0228
Zoo	**0.8344±0.0245**	0.7927±0.0630	0.7094±0.0756

The average intracluster distance of each cluster and the average intercluster distance between each pair of clusters has been presented in Table 5 to Table 10, where we can see that for Voting, WBDC, and Car datasets the W-P intercluster distance increase in comparison with DM3 results, while Mushroom, Soybean, and Zoo datasets the average W-P intracluster distance decrease in comparison with DM3, this is reasonable because by (4) we can deduce that for datasets with larger number of attributes the weights become small, so the distance between samples is short. Moreover, in Voting and Car datasets the difference between the average intercluster and intracluster distance with W-P metric is greater than DM3 metric results, in the rest of datasets, the difference between the average intercluster and intracluster distance is very similar.

Table 5. Average intracluster/intercluster distance obtained by the different metrics on the voting data set

	DM3			W-P metric	
Clusters	C_1	C_2	Clusters	C_1	C_2
C_1	0.3542	0.6380	C_1	0.2909	0.7295
C_2	0.6380	0.2237	C_2	0.7295	0.3345

Table 6. Average intracluster/intercluster distance obtained by the different metrics on the WBCD data set

	DM3			W-P metric	
Clusters	C_1	C_2	Clusters	C_1	C_2
C_1	0.1699	0.6380	C_1	0.7946	0.8789
C_2	0.6380	0.2655	C_2	0.8789	0.4938

Table 7. Average intracluster/intercluster distance obtained by the different metrics on the mushroom data set

	DM3			W-P metric	
Clusters	C_1	C_2	Clusters	C_1	C_2
C_1	0.3882	0.5774	C_1	0.3578	0.4450
C_2	0.5774	0.3876	C_2	0.4450	0.2643

Table 8. Average intracluster/intercluster distance obtained by the different metrics on the soybean data set

	DM3				W-P metric				
Clusters	C_1	C_2	C_3	C_4	Clusters	C_1	C_2	C_3	C_4
C_1	0.1095	0.6149	0.5233	0.5651	C_1	0.0597	0.2081	0.2339	0.2669
C_2	0.6149	0.0744	0.8877	0.8287	C_2	0.2081	0.0587	0.2081	0.1891
C_3	0.5233	0.8877	0.1392	0.3752	C_3	0.2339	0.2081	0.0618	0.1404
C_4	0.5651	0.8287	0.3752	0.1839	C_4	0.2669	0.1891	0.1404	0.0534

Table 9. Average intracluster/intercluster distance obtained by the different metrics on the car data set

	DM3				W-P metric				
Clusters	C_1	C_2	C_3	C_4	Clusters	C_1	C_2	C_3	C_4
C_1	0.5165	0.5688	0.5753	0.5881	C_1	0.6316	0.7446	0.7474	0.7218
C_2	0.5688	0.4526	0.4526	0.4306	C_2	0.7446	0.5881	0.7445	0.7323
C_3	0.5753	0.4526	0.3721	0.3961	C_3	0.7474	0.7445	0.5571	0.7409
C_4	0.5881	0.4306	0.3961	0.2612	C_4	0.7218	0.7323	0.7409	0.5555

Table 10. Average intracluster/intercluster distance obtained by the different metrics on the zoo data set

	DM3							W-P metric							
Clusters	C_1	C_2	C_3	C_4	C_5	C_6	C_7	Clusters	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_1	0.17	0.73	0.52	0.67	0.57	0.72	0.77	C_1	0.10	0.37	0.53	0.62	0.51	0.29	0.35
C_2	0.73	0.11	0.44	0.55	0.51	0.42	0.48	C_2	0.37	0.08	0.27	0.34	0.41	0.33	0.33
C_3	0.52	0.44	0.23	0.33	0.30	0.51	0.44	C_3	0.53	0.27	0.08	0.44	0.53	0.35	0.47
C_4	0.67	0.55	0.33	0.06	0.34	0.07	0.52	C_4	0.62	0.34	0.44	0.11	0.19	0.56	0.54
C_5	0.57	0.51	0.30	0.34	0.07	0.52	0.41	C_5	0.51	0.41	0.52	0.19	0.08	0.67	0.49
C_6	0.72	0.42	0.51	0.68	0.52	0.12	0.32	C_6	0.29	0.33	0.35	0.56	0.67	0.04	0.47
C_7	0.77	0.48	0.44	0.46	0.41	.32	0.17	C_7	0.35	0.33	0.47	0.54	0.49	0.47	0.11

5. CONCLUSION

In this work, we introduced a new similarity/dissimilarity measure for categorical data based on the feature space structure. This distance metric is a variation of pairing matching but weighted. We call our method: weighted pairing (W-P) based on feature space-structure. The weights are determined for the number of states that each feature has, indicating which attribute contributes more to the cluster's compact structure. The performance of W-P metric was evaluated in terms of intracluster/intercluster distance, CDI, RI, and NMI index into a K-modes algorithm with support-based initialization in the unsupervised learning framework, and we compare with the distance metric (DM3) and H-SBI methods. The obtained results showed a better performance for W-P than DM3 and H-SBI, we demonstrated that this way of computing a distance is effective in recovering the inherent clustering structures from categorical data when such structures exist, and this can be attributed to the fact that our approach is space-structure based.

ACKNOWLEDGEMENTS

This work was developed under the framework of the research project "Desarrollo de una metodología para la identificación de perfiles de los consumidores del servicio público utilizando técnicas de aprendizaje de máquina" (number 2-20-8) funded by Vice-Rectory for Research of Universidad Tecnológica de Pereira.

REFERENCES

[1] A. Agresti, "An introduction to categorical data analysis," John Wiley & Sons, 2018.
[2] C. Shen, et al., "Efficient dual approach to distance metric learning," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 2, pp. 394-406, 2013.
[3] H. H. Bock and E. Diday, "Analysis of symbolic data: exploratory methods for extracting statistical information from complex data," Springer Science & Business Media, 2012.
[4] V. Chandola, et al., "A framework for exploring categorical data," *Proceedings of the 2009 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics*, 2009.
[5] T. Xiong, et al., "DHCC: Divisive hierarchical clustering of categorical data," *Data Mining and Knowledge Discovery*, vol. 24, no. 1, pp. 103-135, 2012.
[6] D. J. Hand, "Principles of data mining," *Drug safety*, vol. 30, no. 7, pp. 621-622, 2007.
[7] D. R. Wilson and T. R. Martinez, "Improved heterogeneous distance functions," *Journal of artificial intelligence research*, vol. 6, pp. 1-34, 1997.
[8] H. Ralambondrainy, "A conceptual version of the K-means algorithm," *Pattern Recognition Letters*, vol. 16, no. 11, pp. 1147-1157, 1995.
[9] A. Ahmad and L. Dey, "A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set," *Pattern Recognition Letters*, vol. 28, no. 1, pp. 110-118, 2007.
[10] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 645-678, 2005.
[11] A. R. Mahdiraji, "Clustering data stream: A survey of algorithms," *International Journal of Knowledge-based and Intelligent Engineering Systems*, vol. 13, no. 2, pp. 39-44, 2009.
[12] L. Kaufman and P. J. Rousseeuw, "Finding groups in data: an introduction to cluster analysis," John Wiley & Sons, vol. 344, 2009.
[13] Z. Huang and M. K. Ng, "A fuzzy k-modes algorithm for clustering categorical data," *IEEE transactions on Fuzzy Systems*, vol. 7, no. 4, pp. 446-452, 1999.
[14] D. W. Goodall, "A new similarity index based on probability," *Biometrics*, vol. 2, no. 4, pp. 882-907, 1966.
[15] S. Boriah, et al., "Similarity measures for categorical data: A comparative evaluation," *Proceedings of the 2008 SIAM international conference on data mining. Society for Industrial and Applied Mathematics*, 2008.
[16] H. Jia, et al., "A new distance metric for unsupervised learning of categorical data," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 5, pp. 1065-1079, 2015.
[17] S. Q. Le and T. B. Ho, "An association-based dissimilarity measure for categorical data," *Pattern Recognition Letters*, vol. 26, no. 16, pp. 2549-2557, 2005.
[18] D. Dheeru and E. K. Taniskidou, "UCI machine learning repository," 2017. Available: https://archive.ics.uci.edu/.
[19] H. Jia, et al., "A new distance metric for unsupervised learning of categorical data," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 5, pp. 1065-1079, 2015.
[20] A. Kumar and S. Kumar, "A Support Based Initialization Algorithm for Categorical Data Clustering," *Journal of Information Technology Research (JITR)*, vol. 11, no. 2, pp. 53-67, 2018.
[21] S. Zhu and L. Xu, "Many-objective fuzzy centroids clustering algorithm for categorical data," *Expert Systems with Applications*, vol. 96, pp. 230-248, 2018.
[22] Z. Sulc and H. Rezankova, "Evaluation of recent similarity measures for categorical data," *Proceedings of the 17th International Conference Applications of Mathematics and Statistics in Economics. Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wroclaw*, 2014.
[23] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data mining and knowledge discovery*, vol. 2, no. 3, pp. 283-304, 1998.
[24] D. Steinley and M. J. Brusco, "A note on the expected value of the Rand index," *British Journal of Mathematical and Statistical Psychology*, vol. 71, no. 2, pp. 287-299, 2018.
[25] A. Strehl, et al., "Impact of similarity measures on web-page clustering," *Workshop on artificial intelligence for web search (AAAI 2000)*, pp. 58-64, 2000.